*Applied Linguistics 2016: 37/6: 849–880* 
doi:10.1093/applin/amu065 Advance Access published on 5 December 2014

# Investigating Criterial Discourse Features across Second Language Development: Lexical Bundles in Rated Learner Essays, CEFR B1, B2 and C1

[1,2,]*YU-HUA CHEN and [2]PAUL BAKER

[1]School of English, University of Nottingham Ningbo China and [2]Department of Linguistics and English Language, Lancaster University
*E-mail: Yu-Hua.Chen@nottingham.edu.cn; p.baker@lancaster.ac.uk

In this study, we investigated criterial discourse features in L2 writing through the use of recurrent word combinations, a.k.a. lexical bundles, taking a corpus-driven and expert-judged approach by examining L2 English data across various proficiency levels from L1 Chinese learners. Proficiency was determined by a robust rating procedure which is often used in high-stakes tests, instead of the traditional approach of utilizing extra-linguistic judgement such as program levels. Expository and argumentative essays produced by learners were rated by experienced raters and then subjected to post-rating statistical analysis. Three sizeable subcorpora, representing the Common European Framework of Reference B1, B2, and C1 levels, were then selected for investigation. After lexical bundles were retrieved and refined, structures and discourse functions were manually annotated. The findings suggest that learner writing at lower levels tends to share more features with conversation, whereas the discourse of more proficient writing is closer to that of academic prose. The implications and limitations of the study will also be discussed.

## INTRODUCTION

In recent decades, many studies have focused on distinctive features across second language development, that is, features which can be used to distinguish adjacent levels. In a meta-study which summarizes such second language research (SLA) well, Wolfe-Quintero *et al.* (1998) compared 39 studies on second language development in writing and over 100 measures which gauge the development of learners at known proficiency levels in terms of fluency, accuracy, and complexity. The general assumption is that the more proficient a learner is, the more fluent, accurate, and complex will be their language. In these studies, however, proficiency is generally conceptualized through various external criteria such as age or school level. Needless to say, the determination of proficiency will significantly affect the

discriminative power of development measures and hence impact on the validity of analysis. Thomas (1994), in a review article, compared 157 studies of second language acquisition and categorized the means for assessing L2 proficiency into four types: impressionistic judgement, institutional status, in-house assessment instrument, and standardized test. Thomas concluded that sometimes target language proficiency is poorly controlled to the extent that 'it limits the generalizability of research results'. Another issue in this SLA tradition is that these studies generally rely on rather small quantities of empirical data, often on the basis of a small number of subjects, which again makes the generalizability of results dubious. In addition, few developmental studies have attempted to extend the attention towards features relating to discourse.

Different from the traditional L2 developmental research described above, a new trend in recent years has been to use candidate responses in language tests in a search for language features that distinguish learner performance across proficiency levels. This new thread of research has led to collaboration between practitioners from the fields of language testing and SLA. Studies with empirical data retrieved from candidate scripts in high-stakes exams generally include discourse features such as coherence and cohesion in their investigation of learner language development. For instance, with the aim of developing a common scale for the assessment of writing in the Cambridge Main Suite, Hawkey and Barker (2004) describe in detail how they adopted intuitive, qualitative, and quantitative methods and grouped their findings into versatile distinguishing features. The features explored included fluency, organization, lexico-grammatical accuracy, vocabulary range, collocations, and so on. Among studies of this type, Kennedy and Thorp's project (2007) is probably the one that has considered aspects of discourse the most thoroughly. Working with IELTS candidates' argumentative essay-writing across several band scores,[1] the researchers looked at a variety of features, such as rhetorical questions, modality items, discourse markers, subordinators, and coordinators. One of their major findings was that compared with candidates who received lower band scores, the more proficient IELTS candidates used lexico-grammatical markers (e.g. *however*), enumerative markers (e.g. *firstly*), and subordinators (e.g. *because*) much less frequently, and they appeared to be closer to native-speaker usage in this respect. With 130 essays in total, containing 35,464 words across three levels in IELTS writing, their findings underpin the argument that there is some linear relationship underlying the acquisition of discourse features in learner language development. Mayor *et al.* (2007) reported a similar investigation which included discourse features in learner writing at different levels, but with a slightly larger data set from IELTS—186 essays totalling 56,154 words. Using the same corpus-driven approach as in the current study, Staples *et al.* (2013) examined idiomaticity through the use of lexical bundles across three proficiency levels in the TOEFL iBT—defined as high, intermediate, and low—with 480 participants contributing 249,417 words in total. Their quantitative analyses show that learners at lower levels used more

bundles overall, including more bundles extracted from the prompts, and yet the functional analysis reveals a very similar use of lexical bundles across proficiency levels.

Although the integration of corpus approach and the use of test-taker data have contributed significantly to L2 developmental writing research, the lack of a common standard for determining learner proficiency still makes it difficult, if not impossible, to generalize across research results. For example, the learner samples investigated in Kennedy and Thorp (2007) and Staples *et al.* (2013) were culled from IELTS and TOEFL exams, respectively. As learner proficiency was determined by test scores in different tests, the comparability of results was therefore limited. The task types investigated were different as well—the former focused only on argumentative essays, while the latter also included integrated writing with additional input from reading and listening materials.

Recently, various studies have started to identify lexical and grammatical 'criterial features' for CEFR, the Common European Framework of Reference (Council of Europe 2001), most notably the *English Profile* project led by researchers from the University of Cambridge (see Hawkins and Buttery 2010; Hawkins and Filipovic 2012). CEFR is arguably one of the most influential frameworks in language education nowadays; however, little research has addressed the aspect of discourse in the form of formulaic language across CEFR levels. Drawing on previous research, the current study investigates criterial features through the use of lexical bundles across learner writing development with proficiency defined on the CEFR scale. First, the learner data in this study were selected from a learner corpus and rated with a robust procedure, which will be described in detail in the next section. Then, by investigating the use of lexical bundles across CEFR-defined proficiency groups, the present study focuses on discourse features from a phraseological perspective, as opposed to lexical or syntactical aspects that have been extensively researched in L2 developmental studies.

Lexical bundles are recurrent continuous word sequences that are retrieved to satisfy specified frequency and dispersion thresholds, for example, occurring at least 20 times per million words in five texts or more. Determined by a frequency-driven approach, the multi-word units derived in this way are found to have customary pragmatic and/or discourse functions that are used and recognized by the speakers of a language within certain contexts (e.g. Biber *et al.* 2004; Cortes 2004; Biber and Barbieri 2007; Hyland 2008). These high frequency sequences largely straddle the boundary between lexis and syntax, functioning as 'basic building blocks of discourse' (Biber *et al.* 2004: 371).

Adopting a structural and functional taxonomy from Biber and his colleagues (Biber *et al.* 1999; Biber *et al.* 2004; Biber and Barbieri 2007), Chen and Baker (2010) compared non-native student academic writing with native peer student writing and published academic prose; they concluded that L2 students tend to overuse certain types of bundles (e.g.

overstating expressions such as *all over the world*) while underusing some expressions that are typical in academic prose (e.g. noun or prepositional bundles such as *the extent to which* or *in the context of*). Ädel and Erman (2012), similarly, compared learner writing with native student writing, and their results show that native speakers used a wider range of different types of lexical bundles. With regard to learners' lexical bundle use under test conditions, as mentioned earlier, the findings of Staples *et al.* (2013) suggest that there is not much difference across proficiency levels in TOEFL iBT writing in terms of the function and degree of fixedness, except for overall frequency.

Starting from a developmental perspective based on the CEFR scale, this study aims to bridge the gap by integrating areas of language testing research and second language developmental studies via the incorporation of a corpus-driven discourse perspective. Different from Staples *et al.* (2013), who rely heavily on quantitative measures to analyze the use of lexical bundles, we focus on qualitative and quantitative analyses of the overall structural and functional patterns of lexical bundle use that can be used to distinguish between CEFR levels. In addition, we strictly control the L1 background and task type, whereas these two variables are not accounted for in Staples *et al.* (2013).

## DATA AND METHODOLOGY

### Corpus data

The learner data used come from the Longman Learner Corpus (LLC), a large computerized collection of documents written by learners of L2 English, mainly comprising essays and exam scripts contributed by language schools, teachers, and students throughout the world between 1990 and 2002. To avoid having to account for the effects of different L1s, only argumentative or expository pieces written by L1 Chinese learners of L2 English were chosen from the corpus. This resulted in the selection of 1,029 essays.

### Determination of CEFR levels

The procedure for standardizing the judgements used in this study originates from the manual for *Relating Language Examinations to the Common European Framework of Reference for Languages* (Council of Europe 2003). Six band levels are distinguished in the CEFR, from entry level A1 to the highest level C2. Holistic scoring is adopted with the use of a rating scale from the manual, which consists of overall descriptors as well as three analytical criteria: range, coherence, and accuracy (ibid.: 187). The process can be divided into six phases and is summarized in Figure 1. Starting with CEFR familiarization training (Phase 1), five members of the Language Testing Research Group at the researchers' university participated in a benchmarking exercise to select

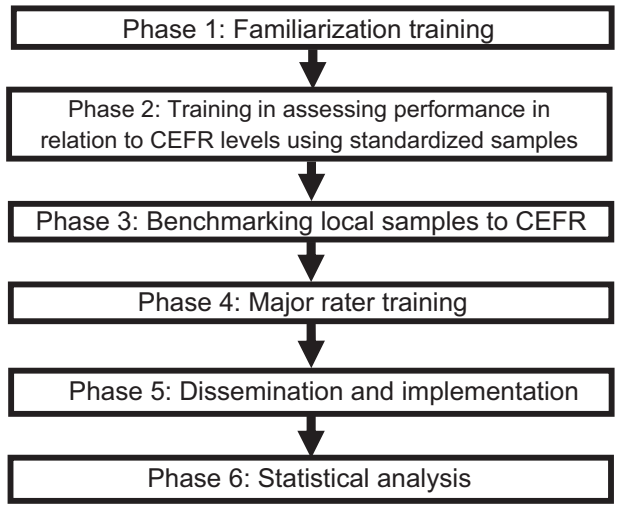


*Figure 1: The process of judgement standardization (extracted and modified from Figure 1.1, Council of Europe 2004)*

appropriate samples from LLC essays for standardization purposes, that is, to select learner essays which were considered representative samples of CEFR levels (Phases 2 and 3). After benchmarking, three experienced raters were trained further on the standardization set of chosen essays (Phase 4). After the three raters passed a post-standardization marking test involving assigning a CEFR level to eight essays, two of the raters independently marked the same set of 1,009 LLC essays, excluding the ones used for training purposes (Phase 5). Any essays which were given different ratings were then sent to the third rater and marked again. Essays therefore received either two or three ratings—two ratings if the first two raters agreed, and three if the first two raters disagreed. All the ratings were then aggregated and subjected to statistical analyses in order to investigate inter-rater reliability, assign a definite CEFR level to each essay, and decide whether each essay would be included in the CEFR-aligned subcorpora or discarded (Phase 6). Inter-rater reliability between the two primary raters was 0.844, while the same index was much lower at 0.766 when the third rater's ratings were included, this being due to the fact that the third rater only marked those essays which received different ratings from the two primary raters. Rasch analysis was also conducted using FACETS (Linacre 2008). In cases of disagreement between raters, essays with a *fit* value higher than 1.3, which suggests erratic rating behavior or atypical learner performance, were excluded.[2]

After the robust rating procedure, three learner subcorpora representing CEFR levels B1, B2, and C1 were established, together forming a 202,154-

*Table 1: Three LLC subcorpora: B1, B2, and C1*

| CEFR Level | Corpus size (word count) | Number of essays | Average essay length |
|---|---|---|---|
| B1 | 26,356 | 189 | 139 |
| B2 | 87,970 | 239 | 368 |
| C1 | 87,828 | 157 | 559 |
| Total | 202,154 | 585 | 345.6 |

word corpus totalling 585 essays (see Table 1). The top C2 level and bottom A1 and A2 levels were discarded because there were insufficient samples. This imbalance in subcorpus size is acknowledged, particularly in the B1 subcorpus where learner writing tends to be substantially shorter. As the number of essays in B1 is still comparable with the other B2 and C1 subcorpora, however, it seems more meaningful to include a broader spectrum of learner language ranging from B1 to C1 rather than spanning only two CEFR levels, B2 and C1. The implications of using a smaller data set for retrieving recurrent word combinations will be discussed in the next section.

## Identification and refinement of lexical bundles

Corpus analysis software, *WordSmith Tools 4.0* (Scott 2004), was used for the automatic retrieval of recurrent word combinations. For comparison with previous research, which has mostly focused on four-word bundles, only the most frequent four-word combinations were investigated. Due to the smaller subcorpus size in this study, it was decided to adopt a dynamic threshold for frequency and dispersion, as discussed in Biber and Barbieri (2007), where lexical bundle use is compared between subcorpora of various sizes ranging from over 1 million words to fewer than 40,000. For the current B2 and C1 subcorpora, lexical bundles are defined as those which occur four times or more in at least three texts, while for the B1 subcorpus the cut-off point is three or more occurrences in at least three texts. A different frequency cut-off was applied because, using a static cut-off point between the three subcorpora with different constituents, for example, occurring four times or more in at least three texts, yielded 86 clusters in the B1 subcorpus but 164 and 169 clusters in the B2 and C1 subcorpora, respectively. A dynamic threshold, on the other hand, leads to an 'optimum' number of clusters in each of the CEFR subcorpora, that is, between 100 and 200 clusters, which is considered to be sufficiently representative and comparable for the subcorpora under examination (cf. Ädel and Erman 2012) and also a

suitable size for manual examination and concordance checks that warrant qualitative analyses. Although the relationship between corpus size, cut-off frequency, and dispersion requires further research, it should be noted that this corpus-driven approach, with a retrieval threshold of around three to five times, has also been reported in several preceding studies which investigated small (sub)corpora, for example, Staples *et al.* (2013) and Biber and Barbieri (2007).

After the automatic retrieval of four-word clusters, two more procedures were performed to filter out context-dependent and overlapping bundles in the data retrieved (Chen and Baker 2010). The former refers to word combinations that recur because of the context in which they are present, such as being part of essay topics or those with proper nouns related to the sociocultural backgrounds of the L2 learners. A few examples of context- or topic-dependent bundles include *the Hong Kong government*, *to the crown court*, and *the countryside is more*. These bundles were manually excluded from the extracted bundle lists as they are not 'building blocks' which display a distinct discourse feature that showcases learner language, as intended by the current study. For example, a large number of the learner essays come from Hong Kong. The reference to *the Hong Kong government* is, therefore, most likely a result of the topics and/or socio-geographical contexts of the learners and thus not considered to be target discourse features in the current investigation. Similarly, *the countryside is more* and *to the crown court* are either part of an essay topic or directly related to it, and hence also discarded. Concordance lines were checked whenever in doubt.[3] The latter—overlapping bundles—refers to four-word lexical bundles which are actually part of a longer expression and yet, as a result of automatic retrieval, the longer expression is split into two or three shorter units. For example, the concordance lines show that *there are a lot* and *are a lot of*, which each occur 11 times, originate from exactly the same 11 contexts. Overlapping word sequences (which were indicative of five-word or even six-word bundles) were manually checked via concordance analyses and combined as appropriate. The above examples of two overlapping bundles were therefore incorporated into a five-word bundle, *there are a lot of*. In the case of partial subsumption (i.e. only some of the concordance lines of two overlapping bundles were identical), a pair of brackets with the character + was added to each combined five-word combination to indicate the extended part of the longer unit. For example, six occurrences of *it is very difficult* and five occurrences of *is very difficult to* share four identical occurrences, hence they were incorporated into *it is very difficult+(to)*. The number of bundles reduced markedly after the two stages of filtering out context-dependent and overlapping instances (Figure 2). Yet, it is believed that the final bundles which were subjected to this scrutiny more genuinely reflect the frequency-related building blocks of discourse in learner language. After the recurrent strings of each proficiency level were finalized, the next step was to categorize the structural and functional associations of

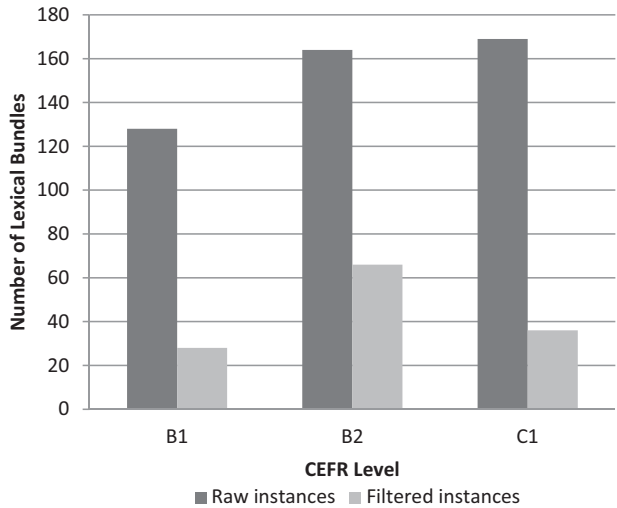


*Figure 2: Number of lexical bundles (types) before and after filtering out context-dependent and overlapping bundles*

each lexical bundle manually, and the results of this are presented in the next section.

## ANALYSES AND RESULTS

### Shared learner bundles

The finalized recurrent strings are presented in Table 2. Five expressions—*on the other hand*, *at the same time*, *for a long time*, *is one of the* and *I would like to*—stand out because they occur in all three levels and also cluster towards the top of the most frequent bundles. Seven bundles are also shared between two adjacent levels, B1 and B2, as well as six shared between B2 and C1, but none are shared between the non-adjacent levels, B1 and C1. Those shared between lower levels (e.g. *a lot of people*, *have a lot of*, *there are so many*) are also notably different from those shared between higher levels (e.g. *it is true that*, *one of the most*, *the end of the*) as the former appears to be more colloquial (e.g. the use of a quantifier such as *a lot of* in four out of six instances) and the latter more formal (e.g. use of the anticipatory *it* structure in two instances and *-of* phrases in three instances). Extensive presence of the quantifier *a lot of* in bundle use is also reported in the register of classroom teaching in Biber *et al.* (2004: 387) but not in the subcorpora of textbooks or academic prose in the same study. The anticipatory *it* structure and prepositional bundles, on the other hand, are found to be characteristic of academic writing (Biber *et al.* 1999; Hyland 2008). The structural and functional differences

*Table 2: Finalized lexical bundles in the LLC B1, B2, and C1 subcorpora and their raw frequencies*

| LLC B1 (27 types) | Frequency | LLC B2 (66 types) | Frequency | LLC C1 (36 types) | Frequency |
|---|---|---|---|---|---|
| **on the other hand** | 10 | **is one of the** | 18 | **on the other hand** | 28 |
| a lot of people[a] | 8 | **on the other hand** | 18 | **at the same time** | 14 |
| I think it is[a] | 8 | **at the same time** | 17 | **is one of the** | 14 |
| if you want to | 8 | a lot of problem(s) | 16 | **I would like to** | 11 |
| have a lot of[a] | 7 | it is (very) difficult+(to)[b] | 12 | it is obvious that+(the) | 11 |
| **at the same time** | 6 | **for a long time** | 11 | one of the most[b] | 11 |
| is very important for | 5 | there are a lot of[a] | 11 | as well as the | 7 |
| **for a long time** | 4 | a lot of people[a] | 10 | it is believed that | 7 |
| I hope I can | 4 | a lot of time | 9 | **for a long time** | 6 |
| **I would like to** | 4 | have/has a lot of[a] | 8 | in the process of | 6 |
| is one of my | 4 | **I would like to** | 7 | it is true that[b] | 6 |
| **is one of the** | 4 | it is also a | 7 | the end of the[b] | 6 |
| more and more people | 4 | most of them are | 7 | the quality of the | 6 |
| there are a lot of[a] | 4 | the most important thing+(is) | 7 | the rest of the | 6 |
| there are so many[a] | 4 | and a lot of | 6 | we can see that | 6 |
| there will be a[a] | 4 | become more and more[a] | 6 | a great deal of | 5 |
| with a lot of | 4 | will not be able to | 6 | all over the world[b] | 5 |
| are more and more | 3 | with the development of | 6 | as a matter of+(fact)[b] | 5 |
| become more and more[a] | 3 | (from)+my point of view | 5 | at the beginning of+(the) | 5 |
| I think the most | 3 | (is)+the best way to | 5 | in order to make | 5 |
| I think this is | 3 | a great number of | 5 | in such a way+(that) | 5 |
| if you don't know | 3 | all over the world[b] | 5 | is a kind of | 5 |

(Continued)

*Table 2: Continued.*

| LLC B1 (27 types) | Frequency | LLC B2 (66 types) | Frequency | LLC C1 (36 types) | Frequency |
|---|---|---|---|---|---|
| it is because the | 3 | is based on the | 5 | we can say that | 5 |
| it is very important | 3 | is very important to | 5 | as a result of | 4 |
| that it is more | 3 | most of the people | 5 | as far as the | 4 |
| the reason is that | 3 | one of the most[b] | 5 | can be divided into | 4 |
| there are many people | 3 | the main reason is | 5 | how to deal with | 4 |
| | | the result of this | 5 | it is hard to | 4 |
| | | there are quite a+(lot of) | 5 | it is not easy+(for) | 4 |
| | | there are too many | 5 | it is very difficult[b] | 4 |
| | | want to be a | 5 | necessary for us to | 4 |
| | | a large amount of | 4 | on the basis of | 4 |
| | | a very important role | 4 | some of them are | 4 |
| | | all of them are | 4 | the relationship between the | 4 |
| | | and to be a | 4 | there are still some | 4 |
| | | are not allowed to | 4 | to cope with the | 4 |
| | | as a matter of fact[b] | 4 | | |
| | | as I have mentioned | 4 | | |
| | | as the result of | 4 | | |
| | | as we all know | 4 | | |
| | | because they are not | 4 | | |
| | | bring a lot of | 4 | | |
| | | but there are still | 4 | | |
| | | him or her to | 4 | | |
| | | I am going to | 4 | | |

(Continued)

*Table 2: Continued.*

| LLC B1 (27 types) | Frequency | LLC B2 (66 types) | Frequency | LLC C1 (36 types) | Frequency |
|---|---|---|---|---|---|
| | | I think it is[a] | 4 | | |
| | | I think that this | 4 | | |
| | | if there is a | 4 | | |
| | | in the following paragraphs | 4 | | |
| | | is more important than | 4 | | |
| | | is the most important | 4 | | |
| | | is totally different from | 4 | | |
| | | it is a good | 4 | | |
| | | it is a very | 4 | | |
| | | it is not a | 4 | | |
| | | it is true that[b] | 4 | | |
| | | should learn how to | 4 | | |
| | | some of them are | 4 | | |
| | | some people think that+(the) | 4 | | |
| | | the end of the[b] | 4 | | |
| | | the quality of the | 4 | | |
| | | the rest of the world | 4 | | |
| | | the result of the | 4 | | |
| | | there are so many[a] | 4 | | |
| | | there will be a[a] | 4 | | |
| | | we can see the | 4 | | |

Bundles occurring in all three levels are indicated in boldface.
[a] indicates bundles occurring in B1 and B2.
[b] indicates bundles occurring in B2 and C1.

*Table 3: Shared learner bundles with normalized frequency per 10,000 words in comparison with other studies*

| Subcorpus (word count) | Present study | | | Staples *et al.* (2013) | | | Chen and Baker (2010) | | | Ädel and Erman (2012) | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | LLC B1 | LLC B2 | LLC C1 | Low | Inter-mediate | High | NNS-CH | NS-EN | EXPERT | NNS-SW | NS-EN |
| Bundle freq. | (26,356) | (87,970) | (87,828) | (74,430) | (87,338) | (87,649) | (146,872) | (155,781) | (164,742) | (863,207) | (247,453) |
| *on the other hand* | 3.8 | 2.0 | 3.2 | 4.3 | 3.9 | 3.1 | 2.5 | 0.3 | 1.2 | 2.6 | 1.6 |
| *at the same time* | 2.3 | 1.9 | 1.6 | 0.8 | 1.5 | 0.8 | 1.6 | 0.3 | 0.6 | 0.8 | 0.6 |
| *is one of the* | 1.5 | 2.0 | 1.6 | — | — | — | 6.1 | 7.7 | — | 0.5 | 0.4 |
| *for a long time* | 1.5 | 1.3 | 0.7 | — | — | — | — | — | — | — | — |
| *I would like to* | 1.5 | 0.8 | 1.3 | 1.0 | 1.2 | 0.9 | — | — | — | — | — |

of the bundles between levels will be discussed in detail in the following sections.

The frequencies of the five bundles shared by all three subcorpora were cross-checked against three other similar studies using the lexical bundle approach to see to what extent learners' preferences for these bundles were sustained regardless of genre or L1 (Table 3). One of these studies considers similar developmental research using TOEFL iBT test-taker data and was conducted by Staples *et al.* (2013), while the other two are comparative: one by Chen and Baker (2010), in which the L2 academic writing of L1 Chinese learners was compared with native English students' writing and expert writing, and the other by Ädel and Erman (2012), in which L1 Swedish students' academic writing was compared with peer L1 English students' writing. Interestingly, the top two bundles in the current study, *on the other hand* and *at the same time*, are shared across all four studies; *on the other hand* is, consistently, the learners' 'all-time' favourite, with a normalized frequency of 2.0 to 4.3 per 10,000 words in various learner groups, whereas native or expert academic writing, wherever reported, has a much lower frequency range of 0.3–1.6 per 10,000 words. As the current study and Chen and Baker (2010) used similar approaches to bundle extraction (including removing context dependent and overlapping bundles), the frequency differences of these two common bundles between these two studies were tested for statistical significance using Paul Rayson's online log-likelihood calculator on the UCREL website (http://ucrel.lancs.ac.uk/llwizard.html). The results confirm the learners' tendency to overuse *on the other hand* and *at the same time* across all three levels in the current study when compared with native student writing or expert writing as reported in Chen and Baker (2010) ($p<0.01$). An additional bundle shared with Chen and Baker (2010) and Ädel and Erman (2012) is *is one of the*. A different bundle, *I would like to,* is also shared with all three levels of Staples *et al.*'s test-taker data. Note that the composition of corpora and the extraction approach to lexical bundles in the above studies may vary from one to another. However, the comparison shows that some bundles, such as *on the other hand* or *at the same time,* consistently constitute important discourse blocks for learner writing regardless of genre or L1 background.

Another interesting finding is that some of the learners' favourite bundles were actually not used appropriately. Take the most frequently used bundle, *on the other hand*, for example. This expression is generally used to compare two different or opposite facts or points of view. A scrutiny of learner use in the concordance lines suggests that learners at lower levels, B1 and B2, tend to use *on the other hand* as a multi-functional discourse marker to link whatever ideas they have, no matter whether these ideas contrast or not, whereas such inappropriate use is not found in C1. About half of the occurrences in B1 data and one third in B2 are found to be semantically problematic. The following examples illustrate the use of this expression in different levels of performance, and an asterisk indicates potentially problematic instances judged by the researchers. This overused learner expression appears to be typical of learner

writing, regardless of proficiency level, and learners at lower levels tend to use it frequently without fully understanding its meaning.

- The only thing they were taught to do is how to be a good wife and a good mother. They lived completely for their husbands and children. ***On the other hand**, they don't have 'egonism'.(LLC-B1)
- She is a vivacious and cute girl. ***On the other hand,** she studies hard. (LLC-B1)
- Everyone has his or her own life, and doesn't like others to disturb. ***On the other hand**, people become more and more salefish and live in their own world. (LLC-B1)
- Many more students are dedicated to much more money without any work. As a result, gambling is in fashion in all universities. ***On the other hand**, teachers never wanted to be a teacher either. They want to be a manager, to get more money with less work... (LLC-B2)
- Though we may hire interpreters, it is not convenient. You can't communicate with them directly at all. ***On the other hand**, we know more and more things about the world when we are working. (LLC-B2)
- As Cantonese is my first language, I acquire it naturally. English, ***on the other hand**, is my second language that I have learnt for nearly two decades. (LLC-B2)
- On one hand, they could not give up their pride in their original identity. ***On the other hand**, their original identity made them feel inferior. (LLC-C1)

Aside from the above bundles, which are characteristic of learner writing, the majority of the lower level B1 bundles differ significantly from those of more advanced C1 writing in terms of both structural and functional associations. The similarities and differences across CEFR levels will be discussed in the remainder of this section.

## Structural characteristics

The structural categorization adopted here follows the taxonomy in the Longman Grammar of Spoken and Written English (LSWE) (Biber *et al.* 1999: 996–1023), where the overall pattern of lexical bundle use is found to be significantly different between academic prose and conversation. In Biber *et al.* (1999: 996), nearly one-third of the bundles in academic prose are noun phrases with *-of* fragments (NP-based, e.g. *the end of the*), and another one third are prepositional phrases with *-of* fragments (PP-based, e.g. *as a result of*). The remaining one third are constructions with a verb component, such as anticipatory *it* patterns and *to*-clause fragments. In comparison, the majority of conversational bundles contain a verb phrase (hence VP-based), and the largest category is 'personal pronoun+verb phrase' (e.g. *I don't know what*). The percentages of each major structural association in the original LSWE

*Table 4: Proportional distribution of lexical bundles (types) across the structural categories in LSWE, B1, B2, and C1 subcorpora (cf. Biber et al. 1999, p. 996)*

| Patterns | | | LSWE | | LLC | | | Example |
|---|---|---|---|---|---|---|---|---|
| | | | CONV | ACAD | B1 | B2 | C1 | |
| | More widely used in 'academic prose' | | | | | | | |
| NP-based | 1 | noun phrase expressions | 4% | 30% | 7% [2][a] | 21% [14] | 17% [6] | *the end of the* |
| PP-based | 2 | prepositional phrase expressions | 3% | 33% | 11% [3] | 14% [9] | 28% [10] | *as a result of* |
| VP-based | 3 | anticipatory *it* + VP/adjectiveP + (complement-clause) | — | 9% | 4% [1] | 3% [2] | 17% [6] | *it is difficult to* |
| | 4 | passive verb + PP fragment | — | 6% | — [0] | 2% [1] | 3% [1] | *is based on the* |
| | 5 | (VP +) *that*-clause fragment | 1% | 5% | 4% [1] | — [0] | — [0] | *that there is a* |
| | More widely used in 'conversation' | | | | | | | |
| | 6 | personal pronoun + verb phrase (+ complement clause) | 44% | — | 19% [5] | 8% [5] | 8% [3] | *I would like to* |
| | 7 | (NP +) copula *be* + NP/adjectiveP | 8% | 2% | 37% [10] | 30% [20] | 11% [4] | *is one of the* |
| | 8 | VP with active verb | 13% | — | 15% [4] | 11% [7] | — [0] | *has a number of* |
| | 9 | *yes-no* and *wh*-question fragment | 12% | — | — [0] | — [0] | — [0] | *can I have a* |
| | 10 | (verb +) *wh*-clause fragment | 4% | — | — [0] | — [0] | — [0] | *know what I mean* |

(Continued)

*Table 4: Continued.*

| Patterns | | LSWE | | LLC | | | Example |
|---|---|---|---|---|---|---|---|
| | | CONV | ACAD | B1 | B2 | C1 | |
| | Patterns in both registers | | | | | | |
| 11 | (verb/adjective +) *to*-clause fragment | 5% | 9% | — [0] | 11% [7] | 6% [2] | *are likely to be* |
| | Subtotal (VP-based) | 87% | 31% | 78% [22] | 64% [42] | 44% [16] | |
| 12 | other expressions | 6% | 6% | 4% [1] | 2% [1] | 11% [4] | *as well as the* |
| | Total | 100% | 100% | 100% [27] | 100% [66] | 100% [36] | |

[a]Raw frequencies are shown in square brackets.

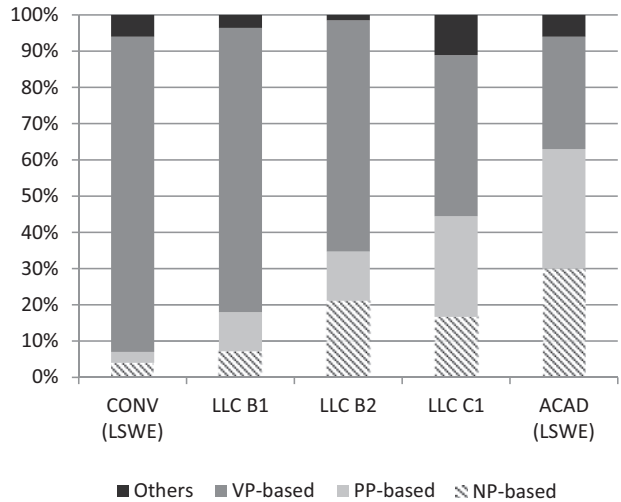


*Figure 3: Distribution of NP-, PP-, and VP-based bundles (types) in LLC B1, B2, and C1 subcorpora in comparison with conversation and academic prose in LSWE*

corpus and the three CEFR subcorpora in the present study are shown in Table 4 and Figure 3. As can be seen, the lowest level, B1, appears to have the highest proportion of VP-based bundles (78%) among the three CEFR groups, and the lowest proportions of NP- and PP-based bundles, thereby being closest to the register of conversation. In contrast, the highest level, C1, shows an opposite pattern, with the lowest proportion of VP-based bundles (44%) and the highest combined proportion of NP- and PP-based bundles (45% in total), thereby being closest to the norm of academic prose.

## NP- and PP-based bundles

If we look further into the subcategories of each structural group, more differences can be identified between CEFR levels. For example, while the majority of NP-based bundles in C1 are noun phrases with *-of* fragments, similar to the pattern of academic prose, a significant proportion of B1 and B2 NP-based bundles fall into the subcategory of 'pre-modifier+noun', for example, *a lot of problem(s)*[4] and *more and more people*, which are not found in the C1 writing. In terms of PP-based bundles, all the B1 bundles and over half of the B2 and C1 bundles in this category are adverbial phrases without *-of* fragments, for example, *on the other hand*, *at the same time*, *all over the world* or *for a long time*, and this is rather different from academic prose, where PP-based bundles are primarily those embedded with *-of* fragments, for example, *in the case of*, *on the basis of*.

*Table 5: Lexical bundles with 'existential there constructions' and the normalized frequency per 10,000 words*

| | LLC Subcorpus Bundle Freq. | B1 | B2 | C1 |
|---|---|---|---|---|
| there + *be* | *there are a lot of* | 1.5 | 1.3 | — |
| | *there are many people* | 1.1 | — | — |
| | *there are so many* | 1.5 | 0.5 | — |
| | *there are too many* | — | 0.6 | — |
| | *there are quite a (lot of)* | — | 0.57 | — |

## VP-based bundles

With regard to VP-based bundles, the most notable pattern emerging from Table 4 is the prevalence of copula *be* constructions in learner writing, particularly at lower levels, which account for over one-third of the bundles at B1 level and nearly one-third at B2 level. The majority of bundles in this subcategory have constructions in the form of 'existential *there*+copula *be*' (e.g. *there are so many*) and '(impersonal pronoun/noun)+copula *be*' (e.g. *is one of the*, *it is also a*, *most of them are*). Again, this finding for the lower levels conforms to the norm of conversation rather than that of the written register (see the section on Pronoun/noun phrase + *be* in Biber *et al.* 1999: 1005–6).

The construction of 'existential *there*+copula *be*' at lower levels often collocates with the quantifiers *a lot of* and *many*—three out of four bundles with this construction in B1 and four out of six in B2 have this pattern. The variations in these existential *there* bundles in B1 and B2 groups are exemplified in Table 5—none were found at C1 level. The existential structure '*there is/are* + NP' is used to stress the notion of existence (Quirk and Greenbaum 1973: 418). Yet, the immoderate use of this structure as well as the superfluous appearance of copula *be* in writing gives rise to a style that appears both simplistic and verbose. A further examination of the concordance lines indicates that many occurrences of bundles with a '*there is/are* + NP' structure are followed by an incorrect verb form or a clause, as a consequence of learner error. Such errors might be due to similar constructions in Chinese, for example, '有 (yǒu, *there is/are*) +NP', which allows existential有 (yǒu) to precede a verb phrase—see the examples below, with the problematic parts underlined:

- More overseas students study in Australia, ***there are a lot of*** advantages are caused by them. (LLC-B1)
- The blind have no choice to do other kind of job because ***there are too many*** companies refuse to hire them. (LLC-B2)
- Why ***there are so many*** prostitutes exists in our society. I think that is because men don't regard women. (LLC-B2)

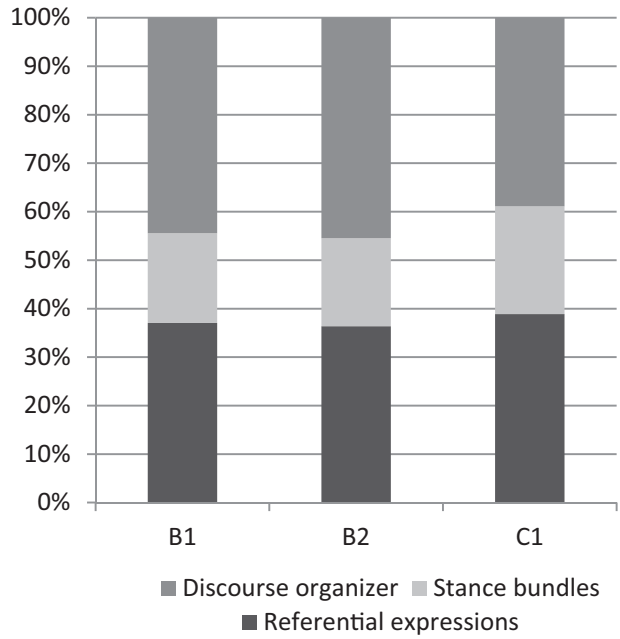


*Figure 4: Functional distribution of lexical bundle types across LLC B1, B2, and C1 subcorpora*

- From the statistic and information, we can see that ***there are too many*** private cars and which cause traffic congestion. (LLC-B2)

## Functional characteristics

Three major discourse functions are distinguished following the taxonomy in Biber *et al.* (2004) and Biber and Barbieri (2007): referential, stance, and discourse organizing. Referential expressions are used to make reference to any entity, including the textual context itself. Stance bundles express the writer's attitude or the certainty of a proposition. Discourse organizers structure prior and coming discourse. Each bundle was manually annotated according to the taxonomy. Concordance lines were checked whenever in doubt, particularly in cases of multi-functionality (i.e. a bundle carries more than one function) and context dependency (i.e. the function of a bundle depends on the context). The rule of thumb when assigning an appropriate function to a lexical bundle is to give priority to 'the most common use' in concordance lines (Biber *et al.* 2004: 384).

As can be seen from the results in Figure 4, there appears to be a very similar distribution of bundle functions across CEFR levels, and a similar pattern is also

*Table 6: Functional categorization of lexical bundles across LLC B1, B2 and C1 subcorpora with normalized frequency per 10,000 words*

| Function | Subfunction | Bundle | B1 | B2 | C1 |
|---|---|---|---|---|---|
| *Referential* | Quantifying | *a great deal of* | — | — | 0.57 |
| | | *a great number of* | — | 0.57 | — |
| | | *a large amount of* | — | 0.45 | — |
| | | *a lot of people* | 3.04 | 1.14 | — |
| | | *a lot of problem(s)* | — | 1.82 | — |
| | | *a lot of time* | — | 1.02 | — |
| | | *all of them are* | — | 0.45 | — |
| | | *and a lot of* | — | 0.68 | — |
| | | *bring a lot of* | — | 0.45 | — |
| | | *have/has a lot of* | 2.66 | 0.91 | — |
| | | *more and more people* | 1.52 | — | — |
| | | *most of the people* | — | 0.57 | — |
| | | *most of them are* | — | 0.80 | — |
| | | *some of them are* | — | 0.45 | 0.46 |
| | | *that it is more* | 1.14 | — | — |
| | | *the rest of the* | — | — | 0.68 |
| | | *the rest of the world* | — | 0.45 | — |
| | | *there are a lot of* | 1.52 | — | — |
| | | *there are many people* | 1.14 | — | — |
| | | *there are quite a (lot of)* | — | 0.57 | — |
| | | *there are so many* | 1.52 | 0.45 | — |
| | | *there are still some* | — | — | 0.46 |
| | | *there are too many* | — | 0.57 | — |
| | | *with a lot of* | 1.52 | — | — |
| | Time/place/ text deixis | *all over the world* | — | 0.57 | 0.57 |
| | | *at the beginning of (the)* | — | — | 0.57 |
| | | *at the same time* | 2.28 | 1.93 | 1.59 |
| | | *for a long time* | 1.52 | 1.25 | 0.68 |
| | | *in the following paragraphs* | — | 0.45 | — |
| | | *the end of the* | — | 0.45 | 0.68 |
| | Framing | *because they are not* | — | 0.45 | — |
| | | *in such a way (that)* | — | — | 0.57 |
| | | *in the process of* | — | — | 0.68 |
| | | *on the basis of* | — | — | 0.46 |

(Continued)

*Table 6: Continued.*

| Function | Subfunction | Bundle | B1 | B2 | C1 |
|---|---|---|---|---|---|
| | | *the main reason is* | — | 0.57 | — |
| | | *the quality of the* | — | 0.45 | 0.68 |
| | | *the reason is that* | 1.14 | — | — |
| | | *the relationship between the* | — | — | 0.46 |
| | | *the result of the* | — | 0.45 | — |
| | | *with the development of* | — | 0.68 | — |
| | | *as a result of* | — | — | 0.46 |
| | | *as the result of* | — | 0.45 | — |
| | | *the result of this* | — | 0.57 | — |
| *Stance* | Epistemic | *as a matter of (fact)* | — | 0.45 | 0.57 |
| | | *as we all know* | — | 0.45 | — |
| | | *become more and more* | 1.14 | 0.68 | — |
| | | *I think it is (very)* | 3.04 | 0.45 | — |
| | | *I think that this* | — | 0.45 | — |
| | | *I think the most* | 1.14 | — | — |
| | | *I think this is* | 1.14 | — | — |
| | | *it is believed that* | — | — | 0.80 |
| | | *it is obvious that (the)* | — | — | 1.25 |
| | | *it is true that* | — | 0.45 | 0.68 |
| | | *some people think that (the)* | — | 0.45 | — |
| | Attitudinal/ modality | *are not allowed to* | — | 0.45 | — |
| | | *I hope I can* | 1.52 | — | — |
| | | *is very important to* | — | 0.57 | — |
| | | *it is (very) difficult (to)* | — | 1.36 | 0.46 |
| | | *it is hard to* | — | — | 0.46 |
| | | *it is not easy (for)* | — | — | 0.46 |
| | | *necessary for us to* | — | — | 0.46 |
| | | *should learn how to* | — | 0.45 | — |
| | | *will not be able to* | — | 0.68 | — |
| *Discourse organizers* | Topic elaboration /clarification | *and to be a* | — | 0.45 | — |
| | | *are more and more* | 1.14 | — | — |
| | | *as well as the* | — | — | 0.80 |
| | | *but there are still* | — | 0.45 | — |
| | | *can be divided into* | — | — | 0.46 |
| | | *how to deal with* | — | — | 0.46 |
| | | *if you don't know* | 1.14 | — | — |

(Continued)

*Table 6: Continued.*

| Function | Subfunction | Bundle | B1 | B2 | C1 |
|---|---|---|---|---|---|
| | | *if you want to* | 3.04 | — | — |
| | | *in order to make* | — | — | 0.57 |
| | | *is a kind of* | — | — | 0.57 |
| | | *is based on the* | — | 0.57 | — |
| | | *is more important than* | — | 0.45 | — |
| | | *is totally different from* | — | 0.45 | — |
| | | *it is a good* | — | 0.45 | — |
| | | *it is a very* | — | 0.45 | — |
| | | *it is also a* | — | 0.80 | — |
| | | *it is because the* | 1.14 | — | — |
| | | *it is not a* | — | 0.45 | — |
| | | *on the other hand* | 3.79 | 2.05 | 3.19 |
| | | *there will be a* | 1.52 | 0.45 | — |
| | | *to cope with the* | — | — | 0.46 |
| | | *want to be a* | — | 0.57 | — |
| | Identification/focus | *(from) my point of view* | — | 0.57 | — |
| | | *(is) the best way to* | — | 0.57 | — |
| | | *a very important role* | — | 0.45 | — |
| | | *as far as the* | — | — | 0.46 |
| | | *as I have mentioned* | — | 0.45 | — |
| | | *him or her to* | — | 0.45 | — |
| | | *is one of my* | 1.52 | — | — |
| | | *is one of the* | 1.52 | 2.05 | 1.59 |
| | | *is the most important* | — | 0.45 | — |
| | | *is very important for* | 1.90 | — | — |
| | | *it is very important* | 1.14 | — | — |
| | | *one of the most* | — | 0.57 | 1.25 |
| | | *the most important thing (is)* | — | 0.80 | — |
| | | *we can say that* | — | — | 0.57 |
| | | *we can see that* | — | — | 0.68 |
| | | *we can see the* | — | 0.45 | — |
| | Topic introduction | *I am going to* | — | 0.45 | — |
| | | *I would like to* | 1.52 | 0.80 | 1.25 |
| | | *if there is a* | — | 0.45 | — |

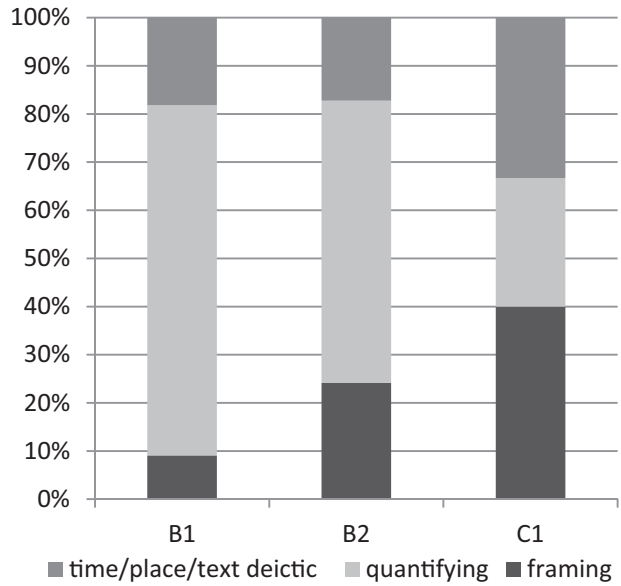


*Figure 5: Distribution of subcategories in referential expressions (types) across LLC B1, B2, and C1 subcorpora*

found in Staples *et al.* (2013), although very few referential expressions are found in their TOEFL iBT data. If we look more closely at the subfunctions in each broad category, however, the distribution is rather different across the levels. Table 6 gives a comprehensive classification of discourse functions, showing the distribution of each bundle across proficiency groups with normalized frequency.

## Referential expressions

Referential expressions can be divided into quantifying, deictic, and framing bundles. Quantifying bundles refer to those that qualify the proposition with expressions related to something potentially gaugeable in terms of size, amount, extent, and so on. One marked difference discovered here is the excessive use of quantifying bundles at lower levels (Figure 5). Many of these contain the informal marker *a lot of* or the determiner *many* (Table 7), which are typical of conversation (Biber *et al.* 1999, 2004). In the present study, these two quantifiers often appear in existential *there* constructions, such as *there are a lot of* or *there are too many*, which signal a very colloquial tone in lower level writing. In comparison, as shown in Figure 5 and Table 7, the number of quantifying bundles at the high level of C1 decreases significantly, and the nature of quantifying bundles also changes to being more 'academic' or 'literate' than conversational (i.e. *a great deal of* and *the rest of the*, both reported

*Table 7: Examples of quantifying bundles in LLC subcorpora*

| Structure | | | Lexical bundle(s) |
|---|---|---|---|
| Quantifier *a lot of* | *a lot of* + noun | B1 | *a lot of people, with a lot of* |
| | | B2 | *a lot of people, a lot of problem(s), a lot of time, and a lot of* |
| | verb + *a lot of* | B1 | *have a lot of* |
| | | B2 | *bring a lot of, has/have a lot of* |
| | *there are* + *a lot of* | B1 | *there are a lot of* |
| | | B2 | *there are a lot of, there are quite a+(lot of)* |
| Other quantifiers | *there are* + *many* | B1 | *there are so many, there are many people* |
| | | B2 | *there are so many, there are too many* |
| | *a/the* + quantifier + *of* | B2 | *a great number of, a large amount of* |
| | | C1 | *a great deal of, the rest of the* |

in native academic writing by Chen and Baker 2010 and Ädel and Erman 2012). The transition from a more colloquial tone to a more literate style as learner language progresses to higher levels may be detectable in the examples below:

- In summer, ***a lot of people*** in the bus and it is crowded. (LLC-B1)
- Everyday ***there are so many*** passengers and goods transport from china mainland to Hong Kong through this way. (LLC-B2)
- ***A great deal of*** attention is paid to the overall presentation, especially to the title page. (LLC-C1)

As for deictic bundles, which make reference to time, place, or text, five out of six bundles in this referential subcategory are adverbial phrases: *at the same time, for a long time, all over the world, at the beginning of* and *in the following paragraphs*.

- Therefore, the people who live in kaohsiung, are quite happy, because there is no wild place to let people visit in kaohsiung ***for a long time***. (LLC-B1)
- I try to judge the identity of customers by their faces and their clothes. ***At the same time***, I listen to their experience, this provides useful information for my future life. (LLC-B2)
- Furthermore, by means of a computer you can have access to all sorts of information, ***all over the world***. (LLC-C1)

Another noticeable pattern that emerges from referential bundles is the subcategory of framing bundles, which is used to specify a particular attribute of

an entity and characteristic of academic writing. Framing bundles account for over one-third of the referential bundles in C1 writing, and many of them are identical or similar to those found in academic prose, as reported in the literature, for example, *in the process of*, *the quality of the* and *on the basis of* (e.g. Biber *et al.* 1999, 2004; Hyland 2008).[5] In comparison, the only one framing bundle in B1 and five out of seven B2 framing bundles are used for inferential purposes to highlight a causal relationship, for example, *the reason is that*, *the result of this/the*. The only inferential bundle in higher level C1 writing is the prepositional phrase *as a result of.* Note that a similar inferential bundle with the definite article *the* is found in B2 writing—*as the result of*. The concordance lines extracted from the subcorpora below indicate that writers in these two proficiency groups used these two variations for the same purpose, and yet the C1 bundle *as a result of* is the one that conforms to the norm in academic prose (Biber *et al.* 1999, 2004; Hyland 2008). It is likely that the distinction between definite and indefinite articles in this case still poses a challenge for B2 learners.

- People in Hong Kong are facing 1997 which is the time when china Government will come and make Hong Kong communist. ***As the result of*** this, many people are immigrating to other countries and the future of Hong Kong is still very difficult to tell. (LLC-B2)
- Such a tendency is partly encouraged by the success of the Guangdong model, and partly ***as a result of*** the weakened control of the central government. (LLC-C1)

## Stance bundles

Stance bundles can be used to convey epistemic or attitudinal/modality senses. Epistemic bundles are used to express the writer's evaluation of a proposition in terms of its certainty or uncertainty (e.g. *as we all know*, *I think this is*). Attitudinal/modality bundles are used to express the writer's attitude, including desire, obligation/directive, prediction, or ability, towards the forthcoming proposition (e.g. *I hope I can*, *it is difficult to*, *will not be able to*). The distribution of subfunctions in stance bundles is presented in Figure 6. Although the dominance of stance bundles across learner levels reported in Staples *et al.* (2013: 220–2) is not found in the current study, a clear shift in the author's voice in stance bundles is identified across the CEFR levels here. At the least proficient level, B1, overt writer visibility is evident in four out of five stance bundles, *I hope I can*, *I think it is (very)*, *I think the most*, and *I think this is*, and the three '*I think*' bundles are all used to express the writer's epistemic evaluation. In contrast, only the anticipatory *it* structure is observed at the higher level of C1 in both epistemic [e.g. *it is not easy* (*for*), *it is* (*very*) *difficult* (*to*)] and attitudinal/modality bundles (e.g. *it is obvious that, it is believed that*). Interestingly, the middle group, B2, shows mixed use of personal and impersonal stance bundles, for example, *as we all know*, *I think it is*, *it is* (*very*) *difficult* (*to*). As can be seen in the following examples, this is another piece of evidence that lower level writing

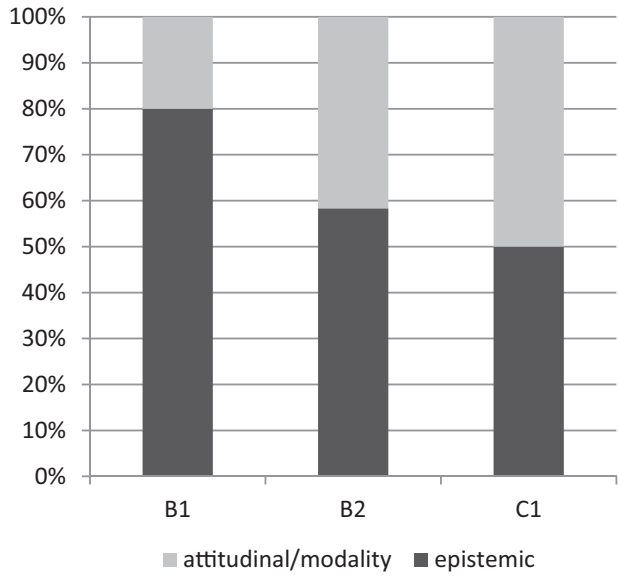


*Figure 6: Distribution of subcategories in stance bundles (types) across LLC B1, B2, and C1 subcorpora*

appears to be more interpersonal and thus more conversational, whereas higher level writing is more impersonal and thus closer to the written register.

- ***I hope I can*** learn English very well, and travel around the England. Making many good memories in my life. (LLC-B1)
- But ***I think it is*** completely wrong, it is the responsibilities of women and men, they are equal.... (LLC-B2)
- ***It is difficult to*** find another country to give them shelter. This put pressure on Hong Kong and also plays an important part in Hong Kong's history. (LLC-B2)
- ***It is believed that*** time and space can affect one's attitude. (LLC-C1)

## Discourse organizers

The final discourse function, discourse organizers, accounts for over one-third of lexical bundles in each of the CEFR subcorpora, and there are three functional subcategories: identification/focus, topic elaboration/clarification, and topic introduction. As can be seen in Figure 7, the distribution of subfunctions shows a similar pattern of use across levels, where topic elaboration/clarification bundles are the largest subcategory across levels but topic introduction bundles are the smallest. In addition to the pervasive explicit discourse marker

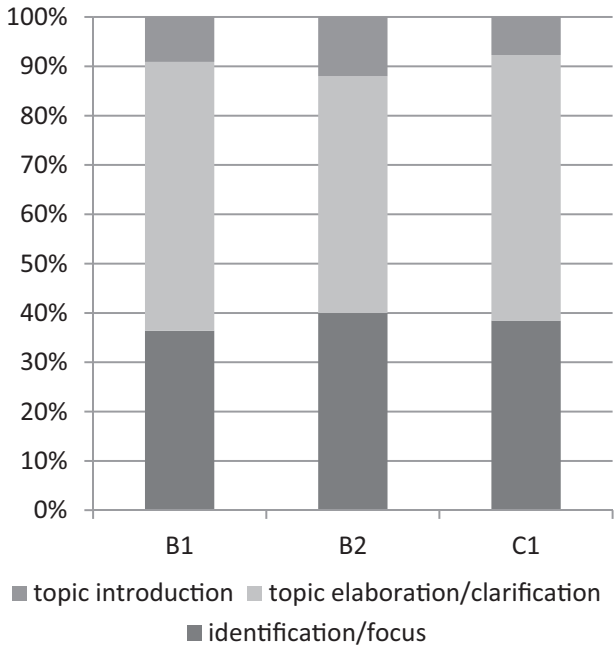


*Figure 7: Distribution of subcategories in discourse organizers (types) across LLC B1, B2, and C1 subcorpora*

*on the other hand* found across all three levels, there was a tendency at the lower level of B1 to use adverbial clausal bundles, including *if you want to* and *if you don't know*, to elaborate or clarify the topic. At the higher level of C1, more proficient writers demonstrated the use of a variety of clausal bundles, such as *in order to make*, *can be divided into*, *how to deal with*, or *to cope with the*. As for the middle group, B2, topic elaboration/clarification bundles are primarily dominated by copula *be* constructions, for example, *it is also a*, *but there are still*, *is more important than*, *is totally different from*, and *it is a very*. Some learner examples of this subfunction can be found below:

- ***If you want to*** buy things, you will very angry that why there is so many people in a shop. (LLC-B1)
- ***It is also a*** very important question that we must answer as university students. (LLC-B2)
- In Gish Jen's In The American Society, the story described a Chinese father who tried to adapt to the American culture ***in order to make*** his family assimilate into the American society... . (LLC-C1)

In terms of identification/focus bundles, all three groups used *is one of the*; variations originating from this bundle are also found at higher levels, such as *the most important thing (is)*, *(is) the best way to* and *is the most important*. One type of common error found across all three proficiency groups in this subcategory is a singular noun following the phrase *one of the* as opposed to the correct plural form. These errors are underlined in the following examples:

- Kenting ***is one of the*** most popular place. (LLC-B1)
- Anyway language ***is one of the*** most important prossession of the human race. (LLC-B2)
- ***The most important thing is*** that advertising encourages competition between manufactures, so keeping prices down and maintaining a high standard. (LLC-B2)
- By the 19th C., China was ***one of the most*** urbanized country in the world. (LLC-C1)

As for topic introduction bundles, there are only three bundle types in this category, and the only common bundle used for this purpose across all three groups is *I would like to*. Yet, note the different uses of this bundle between lower-level and higher-level writing, as in the learner examples below. At B1 and B2 levels, functioning as a fictionalized example, *I would like to* is followed by a material process verb (*change* or *give*) and does not have an explicit discourse-organizing function. In contrast, at C1 level, *I would like to* collocates with a verbal or mental process verb (*comment* or *analyse*), typically associated with academic discourse (for different types of process verbs, see Halliday 1985).

- If I am the teacher in Wen-Tzao junior college, ***I would like to*** change some rules that have been followed since long time ago. (LLC-B1)
- If I have a friend who wants to visit Britain, ***I would like to*** give him some advice or information. (LLC-B2)
- ***I would like to*** comment on two points. (LLC-C1)
- In the following paragraphs ***I would like to*** analyze it from both the demand side and supply side and draw a general conclusion in the end. (LLC-C1)

## DISCUSSION AND CONCLUSION

Drawing on the analyses in the previous section, criterial features in the aspect of discourse across CEFR proficiencies as well as learners' common idiosyncrasies regardless of proficiency have been identified. Lexical bundles in lower-level writing are found to be more verb-heavy (particularly the use of the copula *be*), more personally involved, and to rely more on colloquial quantifiers, including *a lot of* and *many*, hence sharing more features with conversation. In comparison, more proficient writing shows an opposite pattern, having a more impersonal tone with greater use of nominal components in lexical bundles and also sharing more 'academic' or 'literate' bundles with the register of academic prose. Some bundles, however, also appear to persist across all

levels, particularly the omnipresent discourse organizers *on the other hand* and *at the same time*. Although these two expressions are also frequent in academic prose, an overreliance on explicit discourse organizers and the repetitive use of a limited range of familiar formulae are perhaps reasons why non-native writing can still sound awkward, even at more advanced levels.

According to the evidence gathered in the present study, CEFR-B2 is arguably the stage that starts to show signs of transition, whereby learners begin to grasp the distinction between formal and informal writing, as B2 bundles appear to contain as many speech-like elements as written ones. Meanwhile, B1 bundles are highly interactive and conversational, whereas C1 bundles are clearly characterized by a formal style that represents the typical written genre. If we refer back to the CEFR, B2 writers are described as being able to 'make a distinction between formal and informal language with *occasional* less appropriate expressions', and their 'language lacks, however, expressiveness and *idiomaticity* and use of more complex forms is still *stereotypic*' (Council of Europe 2003: 187) [emphasis added]. On the basis of the findings of the present study, the extent of informality discovered in B2 writing is actually greater than simply occasional inappropriacy (e.g. undue use of bundles with the colloquial quantifier *a lot of*). This tendency to be speech-like is, nevertheless, not found in the lexical bundles in C1 writing. In addition, the lack of idiomaticity and the stereotypicality in the use of certain lexical bundles are not only marked in B2 writing but also linger on in C1 writing (e.g. the preference for certain formulae such as *on the other hand, at the same time*). Yet, such features are not seen in the CEFR descriptors at levels above B2. In fact, descriptors of style or formulaicity are rare, except for the one for B2 noted above. Another descriptor which can barely be associated with the discussion here is the statement found in C1: 'The flexibility in style and tone is somewhat limited' (ibid.). As can be seen, the notion of formulaicity and the stylistic aspect disclosed in this study are seldom mentioned in the CEFR scale, yet the evidence suggests that there exist distinctive pragmatic and stylistic developmental features across proficiencies. As most current rating scales generally include lexis, grammar, or coherence as the major definitive criteria for rating, it is therefore recommended to consider adding discourse features other than just cohesion and coherence to the criteria. Moreover, the majority of existing rating scales are constructed on the basis of practitioners' perceptions of typical performance at defined levels, rather than being drawn from learners' actual performance. The CEFR has hence provoked some criticism due to its lack of thorough empirical validation, particularly concerning evidence in the form of learner data (e.g. Alderson 2007; Hulstijn 2007). The findings in this study can thus not only shed light on the discourse aspect of second language development in writing but also provide some empirical underpinning for a large-scale framework of reference for languages, such as the CEFR.

While Staples *et al.* (2013) report that their quantitative analysis reveals a similar pattern of lexical bundle use in terms of functions across levels in their

TOEFL-based study, the mixed approach used in the current study has proved to be effective in identifying criterial discourse features to distinguish between adjacent CEFR levels. It is possible that the proficiency range covered in the current study, from CEFR B1 to C1, is broader than that in Staples *et al.*'s study, in which learner writing was contributed by those who were likely to have prepared for the TOEFL exam. It should also be noted that argumentative and expository essays on a much wider range of topics are investigated in the present study, whereas the learner samples used in Staples *et al.* (2013) are examination responses to two task types (each with two topics only)—independent argumentative essay writing as well as integrated writing, with additional input from reading and listening materials. A closely controlled integrated writing task would probably impact on the production of learner language and thereby the use of lexical bundles. Although the bundles that appeared in the prompts and those clearly related to the topic or task were removed in Staples *et al.*'s study, their functional analysis concluded that 'the majority of bundles are related to the specific topics used in the exam prompts' (ibid.: 222). Examples of such bundles are *according to the lecture/professor/reading*, *the lecture the professor*, and *the second theory is*. Future research could therefore investigate the extent to which task types and the range of essay topics impact on lexical bundle use in learner language.

The limitation inherent in the lexical bundle approach should also be acknowledged here. First of all, lexical bundles represent only one aspect of phraseological competence in learner language. In addition, discursive functions can be expressed by other means such as linking adverbials (e.g. Leedham and Cai 2013). Yet the advantage of using such a corpus-driven approach is that it allows a more systematic and thorough examination of learner language, and any problematic linguistic aspects that might otherwise be implicit can be revealed. The constraint of data size also needs to be discussed. A lexical bundles approach is generally used with native written corpora, which can easily amount to several million words. Conversely, good quality learner data are notoriously difficult to collect. In the case of the current study, learners' L1 background, task type, and proficiency were strictly controlled, which makes it difficult, if not impossible, to gather a data set comparable with native written corpora, although the learner data are already much more substantial than those used in traditional L2 developmental studies. Furthermore, lower-level writing tends to be substantially shorter, and there are usually insufficient samples from the top and bottom proficiency groups. Similar frequency and dispersion thresholds for the lexical bundles approach have, however, been reported in the literature, particularly in studies which looked at lexical bundle use in speech. Through a more detailed examination of lexical bundle structures and functions, the present study has, hopefully, overcome the constraints of learner data size to a large extent.

Finally, it should be stressed that using rated essays to investigate second language development is by no means a circular practice. Performance rating is

a complex judgement process in which a wide range of characteristics can all impact on measurement. In the case of adopting a CEFR rating scale here, the notions of discourse, formulaicity, or idiomaticity are rarely addressed in the assessment criteria grid. The findings in the present study can therefore also be used to flesh out the CEFR descriptors.

*Conflict of interest statement*. None declared.

## NOTES

1 The results of IELTS are reported on a 9-band scale, with 1 being the lowest band score and 9 the highest.
2 According to McNamara (1996, p. 173), the rule-of-thumb acceptable range of a fit value falls within 0.75 to 1.3. For rater consistency, the lower fit figures are generally preferred in the sense that it means the variation between observed and expected values is less than what the model predicts.
3 Given that the source texts of the LLC come from teachers or students who voluntarily contributed their essays, many essays appeared to have come from perhaps a few dozen writing classes. During the initial stage of data selection, efforts were made to avoid including too many essays which responded to identical topics.
4 The bundle *a lot of problem(s)* includes the correct form *a lot of problems* and the erroneous form **a lot of problem*.
5 It has to be noted that the presence of a bundle typical of native/expert writing in learner data does not necessarily guarantee adherence to the native/expert norm, as we have seen in the study. It is, however, not our intention to focus on learner errors if they do not appear as obvious mistakes on the surface. If learners start to use certain expressions, it shows that those expressions are part of learners' language repertoire, even if they are not used correctly.

## REFERENCES

**Ädel, A.** and **B. Erman.** 2012. 'Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach,' *English for Specific Purposes* 31: 81–92.

**Alderson, C.** 2007. 'The CEFR and the need for more research,' *The Modern Language Journal* 91: 659–63.

**Biber, D.** and **F. Barbieri.** 2007. 'Lexical bundles in university spoken and written registers,' *English for Specific Purposes* 26: 263–86.

**Biber, D., S. Conrad,** and **V. Cortes.** 2004. '*If you look at*. . .:Lexical bundles in university teaching and textbooks,' *Applied Linguistics* 25/3: 371–405.

**Biber, D., S. Johansson, G. Leech, S. Conrad,** and **E. Finegan.** 1999. *The Longman Grammar of Spoken and Written English*. Longman.

**Chen, Y.-H.** and **P. Baker.** 2010. 'Lexical bundles in L1 and L2 academic writing,' *Language Learning and Technology* 14/2: 30–49.

**Cortes, V.** 2004. 'Lexical bundles in published and student disciplinary writing: examples from history and biology,' *English for Specific Purposes* 23: 397–423.

**Council of Europe**. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press.

**Council of Europe**. 2003. *Relating Language Examinations to the Common European*

*Framework of Reference for Languages: Learning, Teaching, Assessment. Manual: Preliminary Pilot Version*. DGIV/EDU/LANG 2003, 5. Language Policy Division.

**Council of Europe**. 2004. *Reference Supplement to the Preliminary Pilot Version of the Manual for Relating Language Examinations to the CEFR*. Language Policy Division.

**Halliday, M. A. K.** 1985. *An Introduction to Functional Grammar*. Edward Arnold.

**Hawkey, R.** and **F. Barker.** 2004. 'Developing a common scale for the assessment of writing,' *Assessing Writing* 9/2: 122–59.

**Hawkins, J.** and **P. Buttery.** 2010. 'Criterial features in learner corpora: theory and illustrations,' *English Profile Journal* 1/1: 1–23.

**Hawkins, J.** and **L. Filipovic.** 2012. *Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework (English Profile Studies)*. Cambridge University Press.

**Hulstijn, J. H.** 2007. 'The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency,' *The Modern Language Journal* 91: 663–7.

**Hyland, K.** 2008. '*As can be seen*: Lexical bundles and disciplinary variation,' *English for Specific Purposes* 27/1: 4–21.

**Kennedy, C.** and **D. Thorp.** 2007. 'A corpus investigation of linguistic responses to an IELTS academic writing task' in L. Taylor and P. Falvey (eds): *IELTS Collected Papers: Research in Speaking and Writing Assessment* (Studies in Language Testing 19). Cambridge University Press, pp. 316–78.

**Leedham, M.** and **G. Cai.** 2013. '*Besides... on the other hand*: Using a corpus approach to explore the influence of teaching materials on Chinese students' use of linking adverbials,' *Journal of Second Language Writing* 22/4: 374–89.

**Linacre, J. M.** 2008. *Facets Rasch Measurement Computer Program* (version 3.64.0). Winsteps.com.

**Mayor, B., A. Hewings, S. North, J. Swann,** and **C. Coffin.** 2007. 'A linguistic analysis of Chinese and Greek L1 scripts for IELTS Academic Writing Task 2' in L. Taylor and P. Falvey (eds): *IELTS Collected Papers: Research in Speaking and Writing Assessment* (Studies in Language Testing 19). Cambridge University Press, pp. 250–315.

**McNamara, T. F.** 1996. *Measuring Second Language Performance*. Longman.

**Quirk, R.** and **S. Greenbaum.** 1973. *A University Grammar of English*. Longman.

**Scott, M.** 2004. *WordSmith Tools version 4*. Oxford University Press.

**Staples, S., J. Egber, D. Biber,** and **A. McClair.** 2013. 'Formulaic sequences and EAP development: lexical bundles in the TOEFL iBT writing section,' *English for Specific Purposes* 12: 214–25.

**Thomas, M.** 1994. 'Assessment of L2 proficiency in second language acquisition research,' *Language Learning* 44/2: 307–36.

**Wolfe-Quintero, K., S. Inagaki,** and **H.-Y. Kim.** 1998. *Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity*, Vol. 17. University of Hawaii, Second Language Teaching and Curriculum Center.